RSA: Byzantine-Robust Stochastic Aggregation Methods for Distributed Learning from Heterogeneous Datasets

Liping Li* **Wei Xu*** **Tianyi Chen†** **Georgios B. Giannakis†** **Qing Ling‡**

*Department of Automation, University of Science and Technology of China, Hefei, Anhui, China
†Digital Technology Center, University of Minnesota, Twin Cities, Minneapolis, Minnesota, USA
‡School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong, China

Abstract

This paper proposes a class of robust stochastic subgradient methods for distributed learning from heterogeneous datasets at presence of an unknown number of Byzantine workers. The Byzantine workers, during the learning process, may send arbitrary incorrect messages to the master due to data corruptions, communication failures or malicious attacks, and consequently bias the learned model. The key to the proposed methods is a regularization term incorporated with the objective function so as to robustify the learning task and mitigate the negative effects of Byzantine attacks. The resultant subgradient-based algorithms are termed *Byzantine-Robust Stochastic Aggregation methods*, justifying our acronym RSA used henceforth. In contrast to most of the existing algorithms, RSA does not rely on the assumption that the data are independent and identically distributed (i.i.d.) on the workers, and hence fits for a wider class of applications. Theoretically, we prove that the convergence rate of RSA under Byzantine attacks is the same as that of the stochastic gradient descent method in the Byzantine-free setting and the learning error is determined by the number of Byzantine workers. Numerically, experiments on real dataset corroborate the competitive performance of RSA and a significant complexity reduction compared to the state-of-the-art alternatives.

Introduction

The past decade has witnessed the proliferation of smart phones and Internet-of-Things (IoT) devices. They generate a huge amount of data every day, from which one can learn models of cyber-physical systems and make decisions to improve the welfare of human being. Nevertheless, standard machine learning approaches that require centralizing the training data on one machine or in a datacenter may not be suitable for such applications, as data collected from distributed devices and stored at clouds lead to significant privacy risks (Sicari et al. 2015). To alleviate user privacy concerns, a new distributed machine learning framework called *federated learning* has been proposed by Google and become popular recently (McMahan and Ramage 2017; Smith et al. 2017). Federated learning allows the training data to be kept locally on the owners' devices. Data samples and computation tasks are distributed across multiple workers such as Internet-of-Things (IoT) devices in a smart home, which are programmed to collaboratively learn a model. Parallel implementations of popular machine learning algorithms, such as stochastic gradient descent (SGD), are applied to learning from the distributed data (Bottou ; Dean et al. 2012).

However, federated learning still faces two significant challenges: high communication overhead and serious security risk. While several recent approaches have been developed to tackle the communication bottleneck of distributed learning (Li et al. 2014; Liu et al. 2017; Smith et al. 2017; Chen et al. 2018b), the security issue has not been adequately addressed. In federated learning applications, a number of devices may be highly unreliable or even easily compromised by hackers. We call these devices as Byzantine workers. In this scenario, the learner lacks secure training ability, which makes it vulnerable to failures, not mentioning adversarial attacks (Lynch 1996). For example, SGD, the workhorse of large-scale machine learning, is vulnerable to even one Byzantine worker (Chen, Su, and Xu).

In this context, the present paper studies distributed machine learning under a general Byzantine failure model, where the Byzantine workers can arbitrarily modify the messages transmitted from themselves to the master. With such a model, it simply does not have any constraints on the communication failures or attacks. We aim to develop efficient distributed machine learning methods tailored for this setting with provable performance guarantee.

Related work

Byzantine-robust distributed learning has received increasing attention in recent years. Most of the existing algorithms extend SGD to incorporate the Byzantine-robust setting and assume that the data are independent and identically distributed (i.i.d.) on the workers. Under this assumption, stochastic gradients computed by regular workers are presumably distributed around the true gradient, while those sent from the Byzantine workers to the master could be arbitrary. Thus, the master is able to apply robust estimation techniques to aggregate the stochastic gradients. Typical gradient aggregation rules include geometric median (Chen, Su, and Xu), marginal trimmed mean (Yin et al. 2018a; Xie, Koyejo, and Gupta 2018b), dimensional median (Xie, Koyejo, and Gupta 2018a; Alistarh, Allen-Zhu, and Li

2018), etc. A more sophisticated algorithm termed as Krum selects a stochastic gradient which has minimal local sum of Euclidean distances from a given number of nearest neighbors (Blanchard et al. 2017). Targeting high-dimensional learning, an iterative filtering algorithm is developed in (Su and Xu 2018), which achieves the optimal error rate in the high-dimensional regime. The main disadvantage of these existing algorithms comes from the i.i.d. assumption, which is arguably not the case in federated learning over heterogeneous computing units. Actually, generalizing these algorithms to the non-i.i.d. setting is not straightforward. In addition, some of these algorithms rely on sophisticated gradient selection subroutines, such as those in Krum and geometric median, which incur high computational complexity.

Other related work in this context includes (Yin et al. 2018b) that targets escaping saddle points of nonconvex optimization problems under Byzantine attacks, and (Chen et al. 2018a) that leverages a gradient-coding based algorithm for robust learning. However, the approach in (Chen et al. 2018a) needs to relocate the data points, which is not easy to implement in the federated learning paradigm. Leveraging additional data, (Xie, Koyejo, and Gupta 2018c) studies the trustworthy score-based schemes that guarantee efficient learning even when there is only one non-Byzantine worker, but additional data may not always be available in practice. Our algorithms are also related to robust decentralized optimization studied in, e.g., (Ben-Ameur, Bianchi, and Jakubowicz 2016; Xu, Li, and Ling 2018), which consider optimizing a static or dynamic cost function over a decentralized network with unreliable nodes. In contrast, the focus of this work is Byzantine-robust stochastic optimization.

Our contributions

The contributions of this paper are summarized as follows.

c1) We developed a class of robust stochastic methods abbreviated as RSA for distributed learning over heterogeneous datasets and under Byzantine attacks. RSA has several variants, each tailored for an ℓ_p-norm regularized robustifying objective function.

c2) Performance has been rigorously established for the resultant RSA approaches, in terms of the convergence rate as well as the error caused by the Byzantine attacks.

c3) Extensive numerical tests using the MNIST dataset have been conducted to corroborate the effectiveness of RSA in term of both classification accuracy under Byzantine attacks and runtime.

Distributed SGD

We consider a general distributed system, consisting of a master and m workers, among which q workers are Byzantine (behaving arbitrarily). The goal is to find the optimizer of the following problem:

$$\min_{\tilde{x}\in\mathbb{R}^d} \sum_{i=1}^{m} \mathbb{E}[F(\tilde{x},\xi_i)] + f_0(\tilde{x}). \tag{1}$$

Here $\tilde{x} \in \mathbb{R}^d$ is the optimization variable, $f_0(\tilde{x})$ is a regularization term, and $F(\tilde{x},\xi_i)$ is the loss function of worker i with respect to a random variable ξ_i. Unlike the previous work which assumed the distributed data across the workers are i.i.d., we consider a more practical situation: $\xi_i \sim \mathcal{D}_i$, where $\mathcal{D}_i$ is the data distribution on worker i and could be different to the distributions on other workers.

Algorithm 1 Distributed SGD

Master:

1: Input: $\tilde{x}^0$, α^k. At time $k+1$:
2: Broadcast its current iterate $\tilde{x}^k$ to all workers;
3: Receive all gradients $\nabla F(\tilde{x}^k,\xi_i^k)$ sent by workers;
4: Update the iterate via (2).

Worker i:

1: At time $k+1$:
2: Receive the master's current iterate $\tilde{x}^k$;
3: Compute a local stochastic gradient $\nabla F(\tilde{x}^k,\xi_i^k)$;
4: Send the local stochastic gradient to the server.

In the master-worker architecture, at time $k+1$ of the distributed SGD algorithm, every worker i receives the current model $\tilde{x}^k$ from the master, samples a data point from the distribution ~~$\mathcal{D}_i$~~ with respect to a random variable ξ_i^k, and computes the gradient of the local empirical loss $\nabla F(\tilde{x}^k,\xi_i^k)$. Note that this sampling process can be easily generalized to the mini-batch setting, in which every worker samples multiple i.i.d. data points and computes the averaged gradient of the local empirical losses. The master collects and aggregates the gradients sent by the workers, and updates the model. Its update at time $k+1$ is:

$$\tilde{x}^{k+1} = \tilde{x}^k - \alpha^{k+1}\left(\nabla f_0(\tilde{x}^k) + \sum_{i=1}^{m} \nabla F(\tilde{x}^k,\xi_i^k)\right) \tag{2}$$

where α^{k+1} is a diminishing learning rate at time $k+1$. The distributed SGD is outlined in Algorithm 1.

SGD is vulnerable to Byzantine attacks. While SGD has well-documented performance in conventional large-scale machine learning settings, its performance will significantly degrade at the presence of Byzantine workers (Chen, Su, and Xu). Suppose that some of the workers are Byzantine, they can report arbitrary messages or strategically send well-designed messages according to the information sent by other workers so as to bias the learning process. Specifically, if worker m is Byzantine, at time $k+1$, it can choose one of two following attacks:

a1) sending $\nabla F(\tilde{x}^k,\xi_m^k) = \infty$; and,
a2) sending $\nabla F(\tilde{x}^k,\xi_m^k) = -\sum_{i=1}^{m-1} \nabla F(\tilde{x}^k,\xi_i^k)$.

In any case, the aggregated gradient $\sum_{i=1}^{m} \nabla F(\tilde{x}^k,\xi_i^k)$ used in the SGD update (2) will be either infinite or null, and thus the learned model $\tilde{x}^k$ will either not converge or converge to some arbitrary values. The operation of SGD under Byzantine attacks is illustrated in Figure 1.

Instead of using the simple averaging in (2), robust gradient aggregation rules have been incorporated with SGD in (Blanchard et al. 2017; Chen, Su, and Xu ; Xie, Koyejo, and Gupta 2018a; Yin et al. 2018b; 2018a; Xie, Koyejo, and Gupta 2018c). However, in the federated learning setting, these aggregation rules become less effective due to the d-

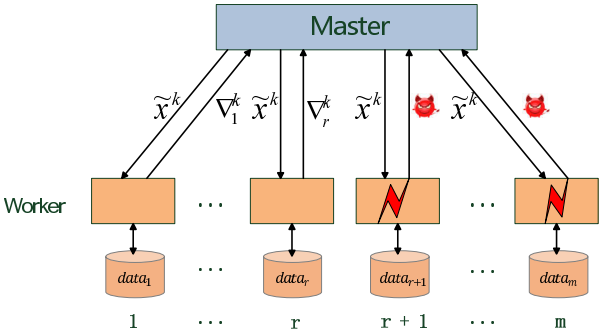


Figure 1: The operation of SGD. There are m workers, r being regular and the rest of $q = m - r$ being Byzantine. The master sends the current iterate to the workers, and the regular workers send back the stochastic gradients. The red devil marks denote the wrong messages that the Byzantine workers send to the master.

ifficulty of distinguishing the statistical heterogeneity from the Byzantine attacks. In what follows, we develop a counterpart of SGD to address the issue of robust learning from distributed heterogeneous data.

RSA for Robust Distributed Learning

Observe that due to the presence of the Byzantine workers, it is meaningless to solve (1), which minimizes the summation of all workers' local expected losses without distinguishing regular and Byzantine workers, because the Byzantine workers can always prevent the learner from accessing their local data and finding the optimal solution. Instead, a less ambitious goal is to find a solution that minimizes the summation of the regular workers' local expected cost functions plus the regularization term:

$$\tilde{x}^* = \arg\min_{\tilde{x}\in\mathbb{R}^d} \sum_{i\in\mathcal{R}} \mathbb{E}[F(\tilde{x},\xi_i)] + f_0(\tilde{x}) \tag{3}$$

Here we denote $\mathcal{B}$ as the set of Byzantine workers and $\mathcal{R}$ as the set of regular workers, with $|\mathcal{B}| = q$ and $|\mathcal{R}| = m - q$. Letting each regular worker i have its local iterate x_i and the master have its local iterate x_0, we obtain an equivalent form to (3):

$$\min_{x:=[x_i;x_0]} \sum_{i\in\mathcal{R}} \mathbb{E}[F(x_i,\xi_i)] + f_0(x_0) \tag{4a}$$

$$\text{s.to} \quad x_0 = x_i,\ \forall i \in \mathcal{R} \tag{4b}$$

where $x := [x_i; x_0] \in \mathbb{R}^{(|\mathcal{R}|+1)\times d}$ is a vector that stacks the regular workers' local variables x_i and the master's variable x_0. The formulation (4) is aligned with the concept of consensus optimization in, e.g., (Shi et al. 2014).

ℓ_1-norm RSA

Directly solving (3) or (4) by iteratively updating $\tilde{x}$ or x is impossible since the identities of Byzantine workers are not available to the master. Therefore, we introduce an ℓ_1-norm regularized form of (4):

$$x^* := \arg\min_{x:=[x_i;x_0]} \sum_{i\in\mathcal{R}} \Big(\mathbb{E}[F(x_i,\xi_i)] + \lambda||x_i - x_0||_1\Big) + f_0(x_0) \tag{5}$$

where λ is a positive constant. The second term in the cost function (5) is the ℓ_1-norm penalty, whose minimization forces every x_i to be close to the master's variable x_0. We will show next that how this relaxed form brings the advantage of robust learning under Byzantine attacks.

In the ideal case that the identities of Byzantine workers are revealed, we can apply a stochastic subgradient method to solve (5). The optimization only involves the regular workers and the master. At time $k+1$, the updates of x_i^{k+1} at regular worker i and x_0^{k+1} at the master are given by:

$$x_i^{k+1} = x_i^k - \alpha^{k+1}\left(\nabla F(x_i^k,\xi_i^k) + \lambda \text{sign}(x_i^k - x_0^k)\right) \tag{6a}$$

$$x_0^{k+1} = x_0^k - \alpha^{k+1}\Big(\nabla f_0(x_0^k) + \lambda\Big(\sum_{i\in\mathcal{R}} \text{sign}(x_0^k - x_i^k)\Big)\Big) \tag{6b}$$

where $\text{sign}(\cdot)$ is the element-wise sign function. Given $a \in \mathbb{R}$, $\text{sign}(a)$ equals to 1 when $a > 0$, -1 when $a < 0$, and an arbitrary value within $[-1, 1]$ when $a = 0$. At time $k+1$, each worker i sends the local iterate x_i^k to the master, instead of sending its local stochastic gradient in distributed SGD. The master aggregates the models sent by the workers to update its own model x_0^{k+1}. In this sense, the updates in (6) are based on model aggregation, different to gradient aggregation in SGD.

Now let us consider how the updates in (6) behave at presence of Byzantine workers. The update of a regular worker i is the same as (6a), which is:

$$x_i^{k+1} = x_i^k - \alpha^{k+1}\left(\nabla F(x_i^k,\xi_i^k) + \lambda \text{sign}(x_i^k - x_0^k)\right). \tag{7}$$

If worker i is Byzantine, instead of sending the value x_i^k computed from (6a) to the master, it sends an arbitrary variable $z_i^k \in \mathbb{R}^d$. The master is unable to distinguish x_i^k sent by a regular worker or z_i^k sent by a Byzantine worker. Therefore, the update of the master at time $k+1$ is no longer (6b), but:

$$x_0^{k+1} = x_0^k - \alpha^{k+1}\Big(\nabla f_0(x_0^k) + \lambda\Big(\sum_{i\in\mathcal{R}} \text{sign}(x_0^k - x_i^k) + \sum_{j\in\mathcal{B}} \text{sign}(x_0^k - z_j^k)\Big)\Big). \tag{8}$$

We term this algorithm as ℓ_1-norm RSA (Byzantine-robust stochastic aggregation).

ℓ_1-norm RSA is robust to Byzantine attacks. ℓ_1-norm RSA is robust to Byzantine attacks due to the introduction of the ℓ_1-norm regularized term to (5). The regularization term allows every x_i to be different from x_0, and the bias is controlled by the parameter λ. This modification robustifies the objective function when any worker is Byzantine and behaves arbitrarily. From the algorithmic perspective, we can observe from the update (8) that the impacts of a regular worker and a Byzantine worker on x_0^{k+1} are similar, no matter the how different the values sent by them to the master are. Therefore, only the number of Byzantine workers

will influence the RSA update (8), rather than the malicious information sent by the Byzantine workers. In this sense, ℓ_1-norm RSA is robust to arbitrary attacks from Byzantine workers. This is in sharp comparison with SGD, which is vulnerable to even a single Byzantine worker.

Generalization to ℓ_p-norm RSA

In addition to solving ℓ_1-norm regularized problem (5), we can also solve the following ℓ_p-norm regularized problem:

$$x^* := \arg\min_{x:=[x_i;x_0]} \sum_{i\in\mathcal{R}} \Big(\mathbb{E}[F(x_i,\xi_i)] + \lambda||x_i - x_0||_p\Big) + f_0(x_0) \tag{9}$$

where $p \geq 1$. Similar to the case of ℓ_1-regularized objective in (5), the ℓ_p norm penalty helps mitigate the negative influence of the Byzantine workers.

Akin to the ℓ_1-norm RSA, the ℓ_p-norm RSA still operates using subgradient recursions. For each regular worker i, its local update at time $k+1$ is:

$$x_i^{k+1} = x_i^k - \alpha^{k+1}\left(\nabla F(x_i^k,\xi_i^k) + \lambda \partial_{x_i}\|x_i^k - x_0^k\|_p\right) \tag{10}$$

where $\partial_{x_i}\|x_i^k - x_0^k\|_p$ is a subgradient of $\|x_i - x_0^k\|_p$ at $x_i = x_i^k$. Likewise, for the master, its update at time $k+1$ is:

$$x_0^{k+1} = x_0^k - \alpha^{k+1}\Big(\nabla f_0(x_0^k) + \lambda\Big(\sum_{i\in\mathcal{R}} \partial_{x_0}\|x_0^k - x_i^k\|_p + \sum_{j\in\mathcal{B}} \partial_{x_0}\|x_0^k - z_j^k\|_p\Big)\Big) \tag{11}$$

where $\partial_{x_0}\|x_0^k - x_i^k\|_p$ and $\partial_{x_0}\|x_0^k - z_j^k\|_p$ are subgradients of $\|x_0 - x_j^k\|_p$ and $\|x_0 - z_j^k\|_p$ at $x_0 = x_0^k$, respectively.

To compute the subgradient involved in ℓ_p-norm RSA, we will rely on the following proposition.

Proposition 1. *Let $p \geq 1$ and b satisfy $\frac{1}{b} + \frac{1}{p} = 1$. For $x \in \mathbb{R}^d$, we have the subdifferential $\partial_x\|x\|_p = \{z \in \mathbb{R}^d : \langle z, x\rangle = \|x\|_p,\ \|z\|_b \leq 1\}$.*

Here and thereafter, we slightly abuse the notation by using ∂ to denote both subgradient and subdifferential. The proof of Proposition 1 is in the supplemental material.

Together with ℓ_1-norm RSA, ℓ_p-norm RSA for robust distributed stochastic optimization under Byzantine attacks is summarized in Algorithm 2 and illustrated in Figure 2.

Remark 1 (Model vs. gradient aggregation)**.** *Most existing Byzantine-robust methods are based on gradient aggregation. Since each worker computes gradient using the same iterate, these methods do not have the consensus issue (Blanchard et al. 2017; Chen, Su, and Xu ; Xie, Koyejo, and Gupta 2018a; Yin et al. 2018b; 2018a; Xie, Koyejo, and Gupta 2018c). However, to enable efficient gradient aggregation, these methods require the data stored in the workers are i.i.d., which is impractical in the federated learning setting. The proposed RSA methods utilize model aggregation, and do not rely on the i.i.d. assumption. On the other hand, the existing gradient aggregation methods generally require to design nontrivial subroutines to aggregate gradients, and hence incur relatively high complexities (Xie, Koyejo, and Gupta 2018a; Blanchard et al. 2017; Su and Xu 2018). In contrast, the proposed RSA methods enjoy much lower complexities, which are the same as that of the standard distributed SGD in the Byzantine-free setting. We shall demonstrate the advantage of RSA in computational time in the numerical tests, in comparison with several state-of-the-art alternatives.*

Algorithm 2 RSA for Robust Distributed Learning

Master:

1: Input: x_0^0, $\lambda > 0$, α^k. At time $k+1$:
2: Broadcast its current iterates x_0^k to all workers;
3: Receive all local iterates x_i^k sent by regular workers or faulty values z_i^k sent by Byzantine workers;
4: Update the iterate via (8) or (11).

Regular Worker i:

1: Input: x_i^0, $\lambda > 0$, α^k. At time $k+1$:
2: Send the current local iterate x_i^k to the master;
3: Receive the master's local iterate x_0^k;
4: Update the local iterate via (7) or (10).

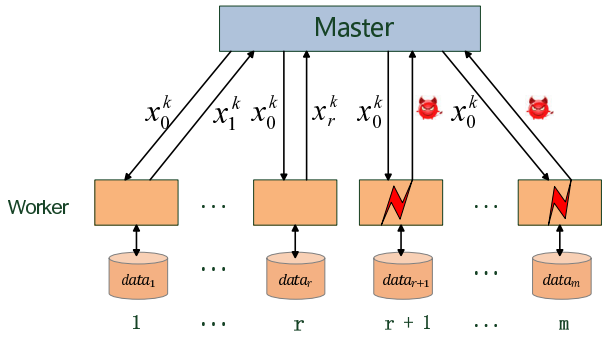


Figure 2: The operation of RSA. There are m workers, r being regular and the rest of $q = m-r$ being Byzantine. The master sends its local variable to the workers, and the regular workers send back their local variables. The red devil marks denote the wrong messages that the Byzantine workers send to the master.

Convergence Analysis

This section analyzes the performance of the proposed RSA methods, with proofs given in the supplementary documents. We make the following assumptions on the cost functions and their gradients.

Assumption 1. (Strong convexity) The local cost functions $\mathbb{E}[F(\tilde{x},\xi_i)]$ and the regularization term $f_0(\tilde{x})$ are strongly convex with constants μ_i and μ_0, respectively.

Assumption 2. (Lipschitz continuous gradients) The local cost functions $\mathbb{E}[F(\tilde{x},\xi_i)]$ and the regularization term $f_0(\tilde{x})$ have Lipschitz continuous gradients with constants L_i and L_0, respectively.

Assumption 3. (Bounded variance) For every worker i, the data sampling is i.i.d. across time such that $\xi_i^k \sim \mathcal{D}_i$. The variance of $\nabla F(\tilde{x},\xi_i)$ is upper bounded by δ_i^2, namely, $\mathbb{E}||\mathbb{E}[\nabla F(\tilde{x},\xi_i)] - \nabla F(\tilde{x},\xi_i)||^2 \leq \delta_i^2$.

Note that Assumptions 1-3 are standard for performance analysis of stochastic gradient-based methods (Nemirovski et al. 2009), and they are satisfied in a wide range of machine learning problems such as ℓ_2-regularized least squares and logistic regression.

We start with investigating the ℓ_p-regularized problem (9), showing the condition under which its optimal solution is consensual and identical to the optimal solution of (3).

Theorem 1. *Suppose that Assumptions 1 and 2 hold. If* $\lambda \geq \lambda_0 := \max_{i\in\mathcal{R}} \|\nabla \mathbb{E}[F(\tilde{x}^*, \xi_i)]\|_b$ *with* $p \geq 1$ *and* b *satisfying* $\frac{1}{b} + \frac{1}{p} = 1$*, then we have* $x^* = [\tilde{x}^*]$*, where* $\tilde{x}^*$ *and* x^* *are the optimal solutions of* (3) *and* (9)*, respectively.*

Theorem 1 asserts that if the penalty constant λ is selected to be large enough, the optimal solution of the regularized problem (9) is the same as that of (3). Next, we shall check the convergence properties of the RSA iterates with respect to the optimal solution of (9) under Byzantine attacks.

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Set the step size of* ℓ_p*-norm RSA* ($p \geq 1$) *as* $\alpha^{k+1} = \min\{\underline{\alpha}, \frac{\overline{\alpha}}{k+1}\}$*, where* $\underline{\alpha}$ *and* $\overline{\alpha}$ *depend on* $\{\mu_0, \mu_i, L_0, L_i\}$*. Then, for* k_0 *satisfying* $\min\{k : \underline{\alpha} \geq \frac{\overline{\alpha}}{k+1}\}$*, we have:*

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq (1-\eta\underline{\alpha})^k \|x^0 - x^*\|^2 + \frac{\underline{\alpha}\Delta_0 + \Delta_2}{\eta},\ k < k_0 \tag{12}$$

and

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \frac{\Delta_1}{k+1} + \overline{\alpha}\Delta_2,\ k \geq k_0 \tag{13}$$

where η*,* Δ_1 *and* $\Delta_2 = \mathcal{O}(\lambda^2 q^2)$ *are certain positive constants.*

Theorem 2 shows that the sequence of local iterates converge sublinearly to the near-optimal solution of the regularized problem (5) or (9). The asymptotic sub-optimality gap is quadratically dependent on the number of Byzantine workers q. Building upon Theorems 1 and 2, we can arrive at the following theorem.

Theorem 3. *Under the same assumptions as those in Theorem 2, if we choose* $\lambda \geq \lambda_0$ *according to Theorem 1, then for a sufficiently large* $k \geq k_0$*, we have:*

$$\mathbb{E}||x^k - [\tilde{x}^*]||^2 \leq \frac{\Delta_1}{k+1} + \overline{\alpha}\Delta_2. \tag{14}$$

If we choose $0 < \lambda < \lambda_0$*, and suppose that the difference between the optimizer of* (9) *and that of* (3) *is bounded by* $||x^* - [\tilde{x}^*]||^2 \leq \Delta_3$*, then for* $k \geq k_0$ *we have:*

$$\mathbb{E}||x^k - [\tilde{x}^*]||^2 \leq \frac{2\Delta_1}{k+1} + 2\overline{\alpha}\Delta_2 + 2\Delta_3. \tag{15}$$

Theorem 3 implies that the sequence of local iterates also converge sublinearly to the near-optimal solution of the original (3). Under a properly selected λ, the sub-optimality gap in the limit is proportional to the number of Byzantine workers. Note that since the $\mathcal{O}(1/k)$ step size is quite sensitive to its initial value (Nemirovski et al. 2009), we use the $\mathcal{O}(1/\sqrt{k})$ step size in our numerical tests. Its corresponding convergence analysis is given in the supplementary document. Regarding the optimal selection of the penalty constant λ and the ℓ_p norm, a remark follows next.

Remark 2 (Optimal selection of λ and p)**.** *Selecting different penalty constant* λ *and* ℓ_p *norms in RSA generally leads to distinct performance. For a fixed* λ*, if a norm* ℓ_p *with a small* p *is used, the dual norm* ℓ_b *has a large* b *and thus results in a small* λ_0 *in Theorem 1. Therefore, the local solutions are likely to be consensual. From the numerical tests, RSA with* ℓ_∞ *norm does not provide competitive performance, while those with* ℓ_1 *and* ℓ_2 *norms work well. On the other hand, for a fixed* p*, a small* λ *cannot guarantee consensus among local solutions, but it gives a small sub-optimality gap* Δ_2*. We recommend to use a* λ *that is relatively smaller than* λ_0*, slightly sacrificing consensus but reducing the sub-optimality gap.*

Numerical Tests

In this section, we evaluate the robustness of the proposed RSA methods to Byzantine attacks and compare them with several benchmark algorithms. We conduct experiments on the MNIST dataset, which has 60k training samples and 10k testing samples, and use softmax regression with an ℓ_2 regularization term $f_0(\tilde{x}) = \frac{0.01}{2}\|\tilde{x}\|^2$. We launch 20 worker processes and 1 master process on a computer with Intel i7-6700 CPU @ 3.40GHz. In the i.i.d. case, the training samples are randomly evenly assigned to the workers. In the heterogeneous case, every two workers evenly share the training samples of every digit. At every iteration, every regular worker estimates its local gradient on a mini-batch of 32 samples. The top-1 accuracy (evaluated with x_0 in RSA and $\tilde{x}$ in the benchmark algorithms) on the test dataset is used as the performance metric.

Benchmark algorithms

We use the SGD iteration (2) without attacks as the oracle, which is referred as **Ideal SGD**. Note that this method is not affected by q, the number of Byzantine workers. The other benchmark algorithms implement the following stochastic gradient aggregation recursion:

$$\tilde{x}^{k+1} = \tilde{x}^k - \alpha^{k+1}\tilde{\nabla}(\tilde{x}^k) \tag{16}$$

where $\tilde{\nabla}(\tilde{x}^k)$ is an algorithm-dependent aggregated stochastic gradient that approximates the gradient direction, at the point $\tilde{x}^k$ sent by the master to the workers. Let the message sent by worker i to the master be v_i^k, which is a stochastic gradient $\nabla F(\tilde{x}^k, \xi_i^k)$ if i is regular, while arbitrary if i is Byzantine. The benchmark algorithms use different rules to calculate the aggregated stochastic gradient.

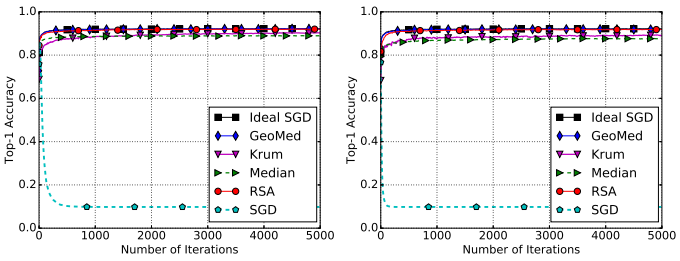
GeoMed (Chen, Su, and Xu). The geometric median of $\{v_i^k : i \in [m]\}$ is denoted by:

$$\text{GeoMed}(\{v_i^k\}) = \arg\min_{v\in\mathbb{R}^d} \sum_{i=1}^m ||v - v_i^k||_2. \tag{17}$$

We use a fast Weiszfeld's algorithm (Weiszfeld and Plastria 2009) to compute the geometric median in the experiments.

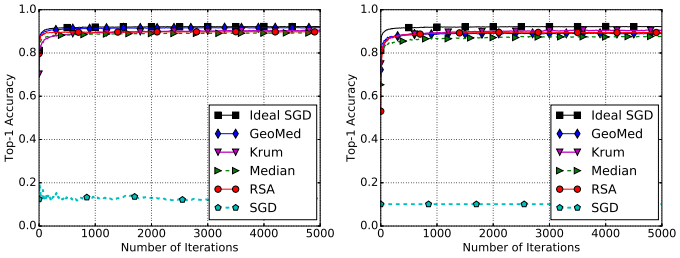
Krum (Blanchard et al. 2017). Krum calculates $\tilde{\nabla}(\tilde{x}^k)$ by:

$$\text{Krum}(\{v_i^k\}) = v_{i^*}^k,\ i^* = \arg\min_{i\in[m]} \sum_{i\to j} ||v_i^k - v_j^k||^2 \tag{18}$$



(a) (b)

Figure 3: Top-1 accuracy under same-value attacks: (a) $q = 4$ and $c = 100$; (b) $q = 8$ and $c = 100$.



(a) (b)

Figure 4: Top-1 accuracy under sign-flipping attacks: (a) $q = 4$ and $\sigma = -4$; (b) $q = 8$ and $\sigma = -4$.

where $i \to j (i \neq j)$ selects the indexes j of the $m - q - 2$ nearest neighbors of v_i^k in $\{v_j^k : j \in [m]\}$ measured by Euclidean distances. Note that q, the number of Byzantine workers, must be known in advance in Krum.

Median (Xie, Koyejo, and Gupta 2018a). The marginal median aggregation rule returns the element-wise median of the vectors $\{v_i^k : i \in [m]\}$.

SGD (Bottou). The classical SGD aggregates $\{v_i^k : i \in [m]\}$ by returning the mean, and is hence not robust to Byzantine attacks.

In the following experiments, step sizes of the benchmark algorithms are all hand-tuned to the best.

Same-value attacks

The same-value attacks set the message sent by a Byzantine worker i as $v_i^k = c\mathbf{1}$. Here $\mathbf{1} \in \mathbb{R}^d$ is an all-one vector and c is a constant, which we set as 100. We consider two different numbers of Byzantine workers, $q = 4$ and $q = 8$, and demonstrate the performance in Figure 3. ℓ_1-norm RSA chooses the regularization parameter $\lambda = 0.07$ and the step size $\alpha^k = 0.001/\sqrt{k}$. When $q = 4$, SGD fails, while RSA and GeoMed are still close to Ideal SGD and outperform Krum and Median. When q is increased to $q = 8$, Krum and Median perform worse than in $q = 4$, while RSA and GeoMed are almost the same as in $q = 4$.

Sign-flipping attacks

The sign-flipping attacks flip the signs of messages (gradients or local iterates) and enlarge the magnitudes. To be specific, a Byzantine worker i first calculates the true value $\hat{v}_i^k$, and then sends $v_i^k = \sigma \hat{v}_i^k$ to the master, where σ is a negative constant. We test $\sigma = -4$ while set $q = 4$ and $q = 8$, as shown in Figure 4. The parameters are $\lambda = 0.07$ and $\alpha = 0.001/\sqrt{k}$ for $q = 4$, while $\lambda = 0.01$ and $\alpha = 0.0003/\sqrt{k}$ for $q = 8$. Not surprisingly, SGD fails in both cases. GeoMed, Median and ℓ_1-norm RSA show similar performance, and Median is slightly worse than the other Byzantine-robust algorithms.

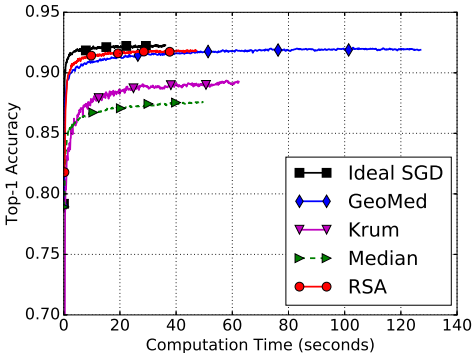


Figure 5: Runtime with $q = 8$ Byzantine workers: same-value attacks with $c = 100$.

Runtime comparison

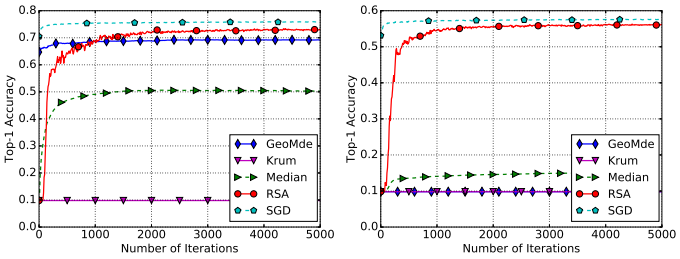
We show in Figure 5 the runtime of the algorithms under the same-value attacks (with parameter $c = 100$) with $q = 8$ Byzantine workers. The total number of iterations for every algorithm is 5000. Though the algorithms are not implemented in a federated learning platform, the comparison clearly demonstrates the additional per-iteration computational costs incurred in handling Byzantine attacks. GeoMed has the largest per-iteration computational cost due to the difficulty of calculating the geometric median. RSA and Median are both slightly slower than Ideal SGD, but faster than Krum. The only computational overhead of RSA than Ideal SDG lies in the computation of sign functions, which is light-weight. Therefore, RSA is advantageous in computational complexity comparing to other complicated gradient aggregation approaches.

Heterogeneous Data

To show the robustness of RSA on heterogeneous dataset, we re-distribute the MNIST data in this way: each two workers associate with the data about the same handwriting digit. In experiment, each Byzantine worker i transmits $v_i^k = v_r^k$, where worker r is one of the regular workers. We set $r = 1$ in the experiment. The results are shown in Figure 6. When $q = 4$, two handwriting numbers' data are not available in the experiment, so the best accuracy is around 0.8. When $q = 10$, the best accuracy is about 0.6. Observe that when $q = 4$, Krum fails, RSA outperforms GeoMed and Median. When q increases to 8, GeoMed , Krum and Median all fail, RSA still perform close to the optimal accuracy.

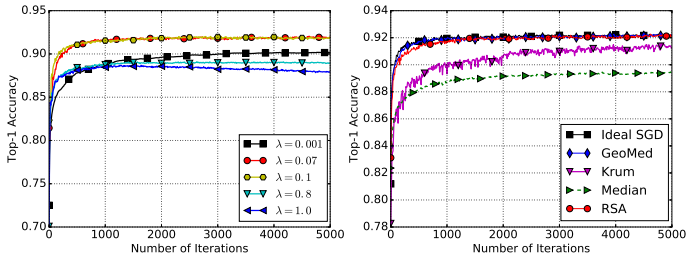
Impact of hyper-parameter λ

We vary the hyper-parameter λ, and show how it affects the performance. We use the same value attacks with $c = 100$,



(a) (b)

Figure 6: Top-1 Accuracy with the attack on heterogeneous data. (a) $q = 4, \lambda = 0.5, \alpha^k = 0.0005/\sqrt{k}$. (b) $q = 8, \lambda = 0.5, \alpha^k = 0.0005/\sqrt{k}$.



(a) (b)

Figure 7: Top-1 Accuracy under (a) same-value attacks with $q = 8$, $c = 100$ and varying λ; (b) no Byzantine attacks.

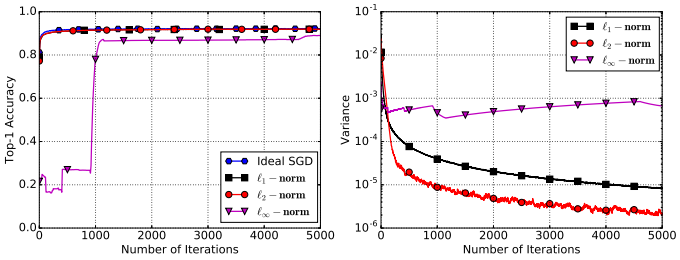
vary λ, run RSA for 5000 iterations, and depict the final top-1 accuracy in Figure 7 (a). The number of Byzantine workers is $q = 8$ and the step sizes are hand-tuned to the best. Observe that when λ is small, a regular worker tends to rely on its own data such that information fusion over the network is slow, which leads to slow convergence and large error. On the other hand, a large λ also incurs remarkable error, as we have investigated in the convergence analysis.

Without Byzantine attacks

In this test, we consider learning without Byzantine workers, and show the performance of all algorithms in Figure 7 (b). ℓ_1-norm RSA chooses the parameter $\lambda = 0.1$ and the step size $\alpha^k = 0.003/\sqrt{k}$. Clearly, RSA and GeoMed are close to Ideal SGD, and significantly outperform Krum and Median. Therefore, robustifying cost in RSA, though introduces bias, does not sacrifice performance in regular case.

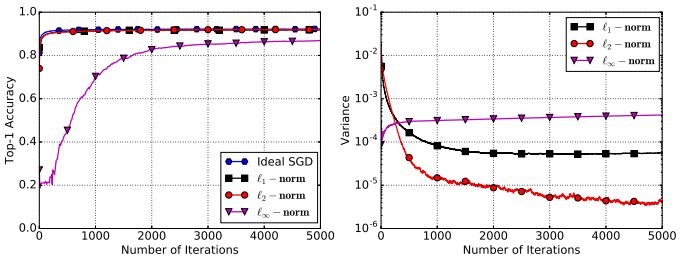
RSA with different norms

Finally, we compare RSA methods regularized with different norms. The results without Byzantine attacks and under the same-value attacks with $q = 8$ and $c = 100$ are demonstrated in Figure 8 and Figure 9, respectively. We consider two performance metrics, top-1 accuracy and variance of the regular workers' local iterates. A small variance means that the regular workers reach a similar solution. The parameters λ and the step size α^k are hand-tuned to the best. Without the Byzantine attacks, ℓ_1 is with $\lambda = 0.1$ and $\alpha^k = 0.001/\sqrt{k}$, ℓ_2 is with $\lambda = 1.4$ and $\alpha^k = 0.001/\sqrt{k}$, while ℓ_∞ is with



(a) (b)

Figure 8: Performance of RSA with different norms, without Byzantine attacks: (a) accuracy; (b) variance of iterates.



(a) (b)

Figure 9: Performance of RSA with different norms, under same-value attacks with $q = 8$ and $c = 100$: (a) top-1 accuracy; (b) variance of local iterates.

$\lambda = 51$ and $\alpha^k = 0.0001/\sqrt{k}$. Under the same-value attacks, ℓ_1 is with $\lambda = 0.07$ and $\alpha^k = 0.001/\sqrt{k}$, ℓ_2 is with $\lambda = 1.2$ and $\alpha^k = 0.001/\sqrt{k}$, while ℓ_∞ is with $\lambda = 20$ and $\alpha^k = 0.0001/\sqrt{k}$. In both cases, ℓ_1-norm RSA and ℓ_2-norm RSA are close in terms of top-1 accuracy, and both of them is better than ℓ_∞-norm RSA. This observation coincides with our convergence analysis, namely, ℓ_∞-norm RSA needs a large λ to ensure consensus, which in turn causes a large error. Indeed, we deliberately choose a not-too-large λ for ℓ_∞-norm RSA so as to reduce the error, but sacrificing the consensus property. Therefore, regarding the variance of the regular workers' local iterates, ℓ_∞-norm RSA is the largest, while ℓ_2-norm RSA is smaller than ℓ_1-norm RSA.

Conclusions

This paper dealt with distributed learning under Byzantine attacks. While existing works mostly focus on the case of i.i.d. data and rely on costly gradient aggregation rules, we developed an efficient variant of SGD for distributed learning from heterogeneous datasets under the Byzantine attacks. The resultant subgradient-based algorithm termed RSA enjoys the sublinear convergence rate of RSA under Byzantine attacks, which is in the same order as SGD in the Byzantine-free setting. Numerically, experiments on real data corroborate the competitive performance of RSA compared to the state-of-the-art alternatives.

References

Alistarh, D.; Allen-Zhu, Z.; and Li, J. 2018. Byzantine stochastic gradient descent. *arXiv preprint arXiv:1803.08917*.

Ben-Ameur, W.; Bianchi, P.; and Jakubowicz, J. 2016. Robust distributed consensus using total variation. *IEEE Trans. Automat. Contr.* 61(6):1550–1564.

Blanchard, P.; Guerraoui, R.; Stainer, J.; and Mhamdi, E. M. E. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 119–129.

Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. In *COMPSTAT'2010*. Heidelberg: Physica-Verlag HD.

Chen, L.; Wang, H.; Charles, Z.; and Papailiopoulos, D. 2018a. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, 902–911.

Chen, T.; Giannakis, G. B.; Sun, T.; and Yin, W. 2018b. LAG: Lazily aggregated gradient for communication-efficient distributed learning. *arXiv preprint:1805.09965*.

Chen, Y.; Su, L.; and Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *In ACM Conference on Measurement and Analysis of Computing Systems*.

Dean, J.; Corrado, G.; Monga, R.; Chen, K.; Devin, M.; Mao, M.; Senior, A.; Tucker, P.; Yang, K.; Le, Q. V.; and Y.Ng, A. 2012. Large scale distributed deep networks. 1223–1231.

Li, M.; Andersen, D. G.; Smola, A. J.; and Yu, K. 2014. Communication efficient distributed machine learning with the parameter server. 19–27.

Liu, Y.; Nowzari, C.; Tian, Z.; and Ling, Q. 2017. Asynchronous periodic event-triggered coordination of multi-agent systems. In *Proc. IEEE Conf. Decision Control*, 6696–6701.

Lynch, N. A. 1996. *Distributed algorithms*. Burlington, MA: Morgan Kaufmann.

McMahan, B., and Ramage, D. 2017. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*.

Nemirovski, A.; Juditsky, A.; Lan, G.; and Shapiro, A. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization* 19(4):1574–1609.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*.

Shi, W.; Ling, Q.; Yuan, K.; Wu, G.; and Yin, W. 2014. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Process.* 62(7):1750–1761.

Sicari, S.; Rizzardi, A.; Grieco, L. A.; and Coen-Porisini, A. 2015. Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks* 76:146–164.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 4427–4437.

Su, L., and Xu, J. 2018. Securing distributed machine learning in high dimensions. *arXiv preprint arXiv:1804.10140*.

Weiszfeld, E., and Plastria, F. 2009. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research* 167(1):7–41.

Xie, C.; Koyejo, O.; and Gupta, I. 2018a. Generalized Byzantine-tolerant SGD. *arXiv preprint arXiv:1802.10116*.

Xie, C.; Koyejo, O.; and Gupta, I. 2018b. Phocas: Dimensional byzantine-resilient stochastic gradient descent. *arXiv preprint arXiv:1805.09682*.

Xie, C.; Koyejo, O.; and Gupta, I. 2018c. Zeno: Byzantine-suspicious stochastic gradient descent. *arXiv preprint arXiv:1805.10032*.

Xu, W.; Li, Z.; and Ling, Q. 2018. Robust decentralized dynamic optimization at presence of malfunctioning agents. *Signal Processing* 153:24–33.

Yin, D.; Chen, Y.; Ramchandran, K.; and Bartlett, P. 2018a. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint:1803.01498*.

Yin, D.; Chen, Y.; Ramchandran, K.; and Bartlett, P. 2018b. Defending against saddle point attack in Byzantine-robust distributed learning. *arXiv preprint arXiv:1806.05358*.

Supplementary Document for "RSA: Byzantine-Robust Stochastic Aggregation Methods for Distributed Learning from Heterogeneous Datasets"

In this supplementary document, we present omitted proofs in the main manuscript.

Proof of Proposition 1

Proof. The proof has two parts.

- Proof of $\{z \in \mathbb{R}^d : \langle z, x\rangle = \|x\|_p,\ \|z\|_b \leq 1\} \subseteq \partial_x \|x\|_p$:

  Considering any $z \in \{z \in \mathbb{R}^d : \langle z, x\rangle = \|x\|_p,\ \|z\|_b \leq 1\}$, we have:

  $$\|x\|_p + \langle z, y - x\rangle = \langle z, y\rangle \leq \|z\|_b \|y\|_p \leq \|y\|_p$$

  where $\langle z, y\rangle \leq \|z\|_b \|y\|_p$ due to Holder's inequality. According to the definition of subdifferential, it holds that $z \in \partial_x \|x\|_p$.

- Proof of $\partial_x \|x\|_p \subseteq \{z \in \mathbb{R}^d : \langle z, x\rangle = \|x\|_p,\ \|z\|_b \leq 1\}$:

  Considering any $z \in \partial_x \|x\|_p$, one can always find a vector x_z that satisfies $\|x_z\|_p = 1$ and $\langle z, x_z\rangle = \|z\|_b$ since $\frac{1}{b} + \frac{1}{p} = 1$. Let $y = \|x\|_p x_z$, we have:

  $$\|y\|_p - \|x\|_p \geq \langle z, y - x\rangle = \langle z, y\rangle - \langle z, x\rangle = \|x\|_p \|z\|_b - \langle z, x\rangle$$

  where the first inequality comes from the definition of subdifferential. Since $\|y\|_p - \|x\|_p = 0$ when $y = \|x\|_p x_z$ and $\|x_z\|_p = 1$, the above result yields $0 \geq \|x\|_p \|z\|_b - \langle z, x\rangle$. However, due to the Holder's inequality it also holds $\|x\|_p \|z\|_b \geq \langle z, x\rangle$. Thus, we must have $\|x\|_p \|z\|_b = \langle z, x\rangle$.

  For $x \neq 0$, we use the definition of subdifferential to derive $\|2x\|_p - \|x\|_p \geq \langle z, x\rangle$ and $\|0\|_p - \|x\|_p \geq \langle z, -x\rangle$, from which we conclude that $\langle z, x\rangle = \|x\|_p$. Since $\|x\|_p \|z\|_b = \langle z, x\rangle$, we have $\|z\|_b = 1$.

  For $x = 0$, it holds that $\langle z, x\rangle = \|x\|_p$. Due to $\|x_z\|_p - \|0\|_p \geq \langle z, x_z\rangle$ from the definition of subdifferential, as well as $\|x_z\|_p = 1$ and $\langle z, x_z\rangle = \|z\|_b$ by hypothesis, we have $\|z\|_b \leq 1$.

Proof of Theorem 1

Proof. Since $\lambda_0 = \max_{i \in \mathcal{R}} \|\nabla \mathbb{E}[F(\tilde{x}^*, \xi_i)]\|_b$ and $\lambda \geq \lambda_0$, we have $\nabla \mathbb{E}[F(\tilde{x}^*, \xi_i)] \in \{\lambda z : \|z\|_b \leq 1\}$. As $p \geq 1$ and $\frac{1}{b} + \frac{1}{p} = 1$, by Proposition 1, we have $\nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] \in \lambda \partial \|0\|_p$ for all $i \in \mathcal{R}$, and consequently:

$$0 \in \nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] + \lambda \partial \|\tilde{x}^* - \tilde{x}^*\|_p,\ \forall i \in \mathcal{R} \tag{19}$$

Also, by $\nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] \in \lambda \partial \|0\|_p$ for all $i \in \mathcal{R}$, there exists $\nu_i \in \partial \|0\|_p$ such that:

$$\nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] + \lambda \nu_i = 0,\ \forall i \in \mathcal{R} \tag{20}$$

Summing (20) up for $i \in \mathcal{R}$, we have:

$$\sum_{i \in \mathcal{R}} \Big(\nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] + \lambda \nu_i \Big) = 0 \tag{21}$$

From the optimality condition of (3), $\sum_{i \in \mathcal{R}} \nabla \mathbb{E}\big[F(\tilde{x}^*, \xi_i)\big] + \nabla f_0(\tilde{x}^*) = 0$. Substituting this equality to (21), we have:

$$\nabla f_0(\tilde{x}^*) - \sum_{i \in \mathcal{R}} \lambda \nu_i = 0 \tag{22}$$

Because every ν_i is a vector satisfying $\nu_i \in \partial \|0\|_p$, it is straightforward to conclude that:

$$0 \in \nabla f_0(\tilde{x}^*) + \sum_{i \in \mathcal{R}} \lambda \partial \|\tilde{x}^* - \tilde{x}^*\|_p \tag{23}$$

Combining (19) and (23), we know that $x^* := [\tilde{x}^*]$ (that is, $x_i^* = \tilde{x}^*$ for all $i \in \mathcal{R}$ and $x_0^* = \tilde{x}^*$) satisfies the optimality condition of (9). This solution is also unique due to Assumption 1.

Proof of Theorem 2

We first give a complete form of Theorem 2 as follows.

A complete form of Theorem 2. Suppose that Assumptions 1, 2 and 3 hold. Set the step size of ℓ_p-norm RSA ($p \geq 1$) as $\alpha^{k+1} = \min\{\underline{\alpha}, \frac{\overline{\alpha}}{k+1}\}$, where $\underline{\alpha} = \min\{\min_{i\in\mathcal{R}} \frac{1}{2(\mu_i+L_i)}, \frac{1}{2(\mu_0+L_0)}\}$, and $\overline{\alpha} > \frac{1}{\eta}$ with $\eta = \min\{\min_{i\in\mathcal{R}} \frac{2\mu_i L_i}{\mu_i+L_i}, \frac{2\mu_0 L_0}{\mu_0+L_0} - \epsilon\}$, and ϵ is any constant within $(0, \frac{2\mu_0 L_0}{\mu_0+L_0})$. Then, there exists a smallest integer k_0 satisfying $\underline{\alpha} \geq \frac{\overline{\alpha}}{k_0+1}$ such that:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq (1-\eta\underline{\alpha})^k \mathbb{E}\|x^0 - x^*\|^2 + \frac{1}{\eta}(\underline{\alpha}\Delta_0 + \Delta_2), \quad \forall k < k_0 \tag{24}$$

and

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \frac{\Delta_1}{k+1} + \overline{\alpha}\Delta_2, \quad \forall k \geq k_0. \tag{25}$$

Here we define

$$\Delta_1 = \max\left\{\frac{\overline{\alpha}^2\Delta_0}{\eta\overline{\alpha}-1}, (k_0+1)\mathbb{E}\|x^{k_0} - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{k_0+1}\right\}$$

as well as

$$\Delta_0 = 16\lambda^2 rd + 16\lambda^2 r^2 d + 2\lambda^2 q^2 d + 2\sum_{i\in\mathcal{R}}\delta_i^2 \quad \text{and} \quad \Delta_2 = \frac{\lambda^2 q^2 d}{\epsilon}.$$

Proof. The proof contains the following steps.

Step 1. From the RSA update (10) at every regular worker i, we have:

$$\begin{aligned}
&\mathbb{E}\|x_i^{k+1} - x_i^*\|^2 \\
&\quad= \mathbb{E}\|x_i^k - x_i^* - \alpha^{k+1}\left(\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\right)\|^2 \\
&\quad= \mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
&\qquad - 2\alpha^{k+1}\mathbb{E}\langle\nabla F(x_i^k, \xi_i^k), x_i^k - x_i^*\rangle - 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle
\end{aligned} \tag{26}$$

Since x_i^k is independent with ξ_i^k, it follows that:

$$\mathbb{E}\big[\langle\nabla F(x_i^k, \xi_i^k), x_i^k - x_i^*\rangle\big] = \mathbb{E}\big[\langle\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big], x_i^k - x_i^*\rangle\big]. \tag{27}$$

Substituting (27) to (26), we have:

$$\begin{aligned}
\mathbb{E}\|x_i^{k+1} - x_i^*\|^2 =& \mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
&- 2\alpha^{k+1}\mathbb{E}\langle\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big], x_i^k - x_i^*\rangle - 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle \\
\overset{(a)}{=}& \mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
&- 2\alpha^{k+1}\mathbb{E}\langle\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big], x_i^k - x_i^*\rangle \\
&- 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p - \lambda\partial_{x_i}\|x_i^* - x_0^k\|_p, x_i^k - x_i^*\rangle
\end{aligned} \tag{28}$$

where in (a), we insert the optimality condition of (9) with respect to x_i, namely, $\mathbb{E}\big[\nabla F(x_i^*, \xi_i)\big] + \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p = 0$, and replace ξ_i by ξ_i^k due to $\xi_i^k \sim \mathcal{D}_i$.

For the second term at the right-hand side of (28), we have:

$$\begin{aligned}
&\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
=&\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p + \nabla F(x_i^k, \xi_i^k) - \nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big]\|^2 \\
\leq&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 + 2\|\nabla F(x_i^k, \xi_i^k) - \nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big]\|^2 \\
\overset{(b)}{\leq}&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 + 2\delta_i^2
\end{aligned} \tag{29}$$

where (b) is due to the bounded variance given by Assumption 3. Plugging the optimality condition $\mathbb{E}\big[\nabla F(x_i^*, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p = 0$ into the right-hand side of (29) yields:

$$\begin{aligned}
&\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
\leq&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p - \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p\|^2 + 2\delta_i^2 \\
\leq&4\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 + 4\lambda^2\mathbb{E}\|\partial_{x_i}\|x_i^k - x_0^k\|_p - \partial_{x_i}\|x_i^* - x_0^*\|_p\|^2 + 2\delta_i^2 \\
\overset{(c)}{\leq}&4\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 + 16\lambda^2 d + 2\delta_i^2
\end{aligned} \tag{30}$$

where (c) holds true because the b-norm of $\partial_{x_i}\|x_i^k - x_0^k\|_p$ (or $\partial_{x_i}\|x_i^* - x_0^*\|_p$) is no larger than 1 according to Proposition 1, and thus the absolute value of every element of the d-dimensional vector $\partial_{x_i}\|x_i^k - x_0^k\|_p$ (or $\partial_{x_i}\|x_i^* - x_0^*\|_p$) is no larger than 1.

For the third term at the right-hand side of (28), since $\mathbb{E}\big[F(x_i, \xi_i^k)\big]$ is strongly convex and has Lipschitz continuous gradients by hypothesis, we have (Nesterov):

$$\mathbb{E}\langle \nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big], x_i^k - x_i^*\rangle \geq \frac{\mu_i L_i}{\mu_i + L_i}\mathbb{E}\|x_i^k - x_i^*\|^2 + \frac{1}{\mu_i + L_i}\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 \tag{31}$$

Substituting (30) and (31) into (28), we have:

$$\begin{aligned}
\mathbb{E}\|x_i^{k+1} - x_i^*\|^2 \leq & \left(1 - \frac{2\alpha^{k+1}\mu_i L_i}{\mu_i + L_i}\right)\mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2(16\lambda^2 d + 2\delta_i^2) \\
& - 2\alpha^{k+1}\left(\frac{1}{\mu_i + L_i} - 2\alpha^{k+1}\right)\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 \\
& - 2\alpha^{k+1}\mathbb{E}\langle \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p - \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p, x_i^k - x_i^*\rangle \\
\leq & \big(1 - \eta\alpha^{k+1}\big)\mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2(16\lambda^2 d + 2\delta_i^2) \\
& - 2\alpha^{k+1}\mathbb{E}\langle \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p - \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p, x_i^k - x_i^*\rangle
\end{aligned} \tag{32}$$

where we drop the term of $\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2$ because $\frac{1}{\mu_i + L_i} - 2\alpha^{k+1} \geq 0$ according to the step size rule.

Step 2. From the RSA update (11) at the master, we have:

$$\begin{aligned}
& \mathbb{E}\|x_0^{k+1} - x_0^*\|^2 \\
=& \mathbb{E}\left\|x_0^k - x_0^* - \alpha^{k+1}\left(\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right)\right\|^2 \\
=& \mathbb{E}\|x_0^k - x_0^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\left\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p v\right\|^2 \\
& - 2\alpha^{k+1}\mathbb{E}\langle \nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p, x_0^k - x_0^*\rangle - 2\alpha^{k+1}\mathbb{E}\langle \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p, x_0^k - x_0^*\rangle.
\end{aligned} \tag{33}$$

For the second term at the right-hand side of (33), we have:

$$\begin{aligned}
& \mathbb{E}\left\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right\|^2 \\
\leq & 2\mathbb{E}\left\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\right\|^2 + 2\lambda^2\mathbb{E}\left\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right\|^2.
\end{aligned} \tag{34}$$

Since every element of the d-dimensional vector $\partial_{x_0}\|x_0^k - z_j^k\|_p$ is within $[-1, 1]$, it holds:

$$\mathbb{E}\left\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right\|^2 \leq q^2 d. \tag{35}$$

For $\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\|^2$, we insert the optimality condition of (9) with respect to x_0, namely, $\nabla f_0(x_0^*) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_i^* - x_0^*\|_p = 0$ to obtain:

$$\begin{aligned}
& \mathbb{E}\left\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\right\|^2 \\
\leq & \mathbb{E}\left\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p - \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_i^* - x_0^*\|_p\right\|^2 \\
\leq & 2\mathbb{E}\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*)\|^2 + 2\lambda^2\mathbb{E}\left\|\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p - \sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^* - x_i^*\|_p\right\|^2 \\
\leq & 2\mathbb{E}\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*)\|^2 + 8\lambda^2 r^2 d
\end{aligned} \tag{36}$$

Substituting (35) and (36) into (34) yields:

$$\begin{aligned}&\mathbb{E}\|\nabla f_0(x_0^k)+\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p+\lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k-z_j^k\|_p\|^2\\ \leq&4\mathbb{E}\|\nabla f_0(x_0^k)-\nabla f_0(x_0^*)\|^2+16\lambda^2r^2d+2\lambda^2q^2d\end{aligned}\tag{37}$$

For the third term at the right-hand side of (33), we insert $\nabla f_0(x_0^*)+\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_i^*-x_0^*\|_p=0$, the optimality condition of (9) with respect to x_0, and obtain:

$$\begin{aligned}&\mathbb{E}\langle\nabla f_0(x_0^k)+\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p,x_0^k-x_0^*\rangle\\ =&\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p-\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^*-x_i^*\|_p,x_0^k-x_0^*\rangle+\mathbb{E}\langle\nabla f_0(x_0^k)-\nabla f_0(x_0^*),x_0^k-x_0^*\rangle\\ \overset{(d)}{\geq}&\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p-\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^*-x_i^*\|_p,x_0^k-x_0^*\rangle+\frac{\mu_0L_0}{\mu_0+L_0}\mathbb{E}\|x_0^k-x_0^*\|^2+\frac{1}{\mu_0+L_0}\mathbb{E}\|\nabla f_0(x_0^k)-\nabla f_0(x_0^*)\|^2\end{aligned}\tag{38}$$

where (d) is due to the fact that f_0 is strongly convex and has Lipschitz continuous gradients (cf. Assumptions 1 and 2).

For the last term at the right-hand side of (33), it holds for any $\epsilon>0$ that:

$$\begin{aligned}2\mathbb{E}\Big\langle\lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k-z_j^k\|_p,x_0^k-x_0^*\Big\rangle\leq&\epsilon\mathbb{E}\|x_0^k-x_0^*\|^2+\frac{\lambda^2}{\epsilon}\mathbb{E}\Big\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k-z_j^k\|_p\Big\|^2\\ \leq&\epsilon\mathbb{E}\|x_0^k-x_0^*\|^2+\frac{\lambda^2q^2d}{\epsilon}.\end{aligned}\tag{39}$$

Substituting (37), (38) and (39) into (33), we have:

$$\begin{aligned}\mathbb{E}\|x_0^{k+1}-x_0^*\|^2\leq&\left(1-(\frac{2\mu_0L_0}{\mu_0+L_0}-\epsilon)\alpha^{k+1}\right)\mathbb{E}\|x_0^{k-1}-x_0^*\|^2+(\alpha^{k+1})^2(16\lambda^2r^2d+2\lambda^2q^2d)+\frac{\alpha^{k+1}\lambda^2q^2d}{\epsilon}\\ &-\alpha^{k+1}\left(\frac{2}{\mu_0+L_0}-4\alpha^{k+1}\right)\mathbb{E}\|\nabla f_0(x_0^k)-\nabla f_0(x_0^*)\|^2\\ &-2\alpha^{k+1}\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p-\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^*-x_i^*\|_p,x_0^k-x_0^*\rangle\\ \leq&(1-\eta\alpha^{k+1})\mathbb{E}\|x_0^k-x_0^*\|^2+(\alpha^{k+1})^2(16\lambda^2r^2d+2\lambda^2q^2d)+\frac{\alpha^{k+1}\lambda^2q^2d}{\epsilon}\\ &-2\alpha^{k+1}\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p-\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^*-x_i^*\|_p,x_0^k-x_0^*\rangle.\end{aligned}\tag{40}$$

We drop the term of $\mathbb{E}\|\nabla f_0(x_0^k)-\nabla f_0(x_0^*)\|^2$ because $\frac{1}{\mu_0+L_0}-2\alpha^{k+1}\geq0$ according to the step size rule.

Step 3. Denote $g_p(x)=\sum_{i\in\mathcal{R}}\|x_i-x_0\|_p$. Since $g_p(x)$ is convex, we have:

$$\begin{aligned}&\langle\partial_xg_p(x^k)-\partial_xg_p(x^*),x^k-x^*\rangle\\ =&\sum_{i\in\mathcal{R}}\langle\partial_{x_i}\|x_i^k-x_0^k\|_p-\partial_{x_i}\|x_i^*-x_0^*\|_p,x_i^k-x_i^*\rangle+\sum_{i\in\mathcal{R}}\langle\partial_{x_0}\|x_0^k-x_i^k\|_p-\partial_{x_0}\|x_0^*-x_i^*\|_p,x_0^k-x_0^*\rangle\geq0.\end{aligned}\tag{41}$$

Summing up (32) for all $i\in\mathcal{R}$, adding (40) and substituting (41), we have:

$$\begin{aligned}\mathbb{E}\|x^{k+1}-x^*\|^2\leq&(1-\eta\alpha^{k+1})\mathbb{E}\|x^k-x^*\|^2+(\alpha^{k+1})^2(16\lambda^2rd+16\lambda^2r^2d+2\lambda^2q^2d+2\sum_{i\in\mathcal{R}}\delta_i^2)+\frac{\alpha^{k+1}\lambda^2q^2d}{\epsilon}\\ =&(1-\eta\alpha^{k+1})\mathbb{E}\|x^k-x^*\|^2+(\alpha^{k+1})^2\Delta_0+\alpha^{k+1}\Delta_2\end{aligned}\tag{42}$$

where for simplicity we denote:

$$\Delta_0=16\lambda^2rd+16\lambda^2r^2d+2\lambda^2q^2d+2\sum_{i\in\mathcal{R}}\delta_i^2\quad\text{and}\quad\Delta_2=\frac{\lambda^2q^2d}{\epsilon}.\tag{43}$$

According to the step size rule $\alpha^{k+1} = \min\{\underline{\alpha}, \frac{\overline{\alpha}}{k+1}\}$, there exists a smallest integer k_0 satisfying $\underline{\alpha} \geq \frac{\overline{\alpha}}{k_0+1}$ such that $\alpha^{k+1} = \underline{\alpha}$ when $k < k_0$ and $\alpha^{k+1} = \frac{\overline{\alpha}}{k+1}$ when $k \geq k_0$. Then for all $k < k_0$, (42) becomes:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq (1 - \eta\underline{\alpha})\mathbb{E}\|x^k - x^*\|^2 + (\underline{\alpha})^2\Delta_0 + \underline{\alpha}\Delta_2, \quad \forall k < k_0. \tag{44}$$

By definitions $\eta = \min\{\min_{i\in\mathcal{R}} \frac{2\mu_i L_i}{\mu_i + L_i}, \frac{2\mu_0 L_0}{\mu_0 + L_0} - \epsilon\}$ and $\underline{\alpha} = \min\{\min_{i\in\mathcal{R}} \frac{1}{2(\mu_i + L_i)}, \frac{1}{2(\mu_0 + L_0)}\}$, $\eta\underline{\alpha} \in (0, \frac{1}{4})$. Applying telescopic cancellation to (44) through time 0 to $k < k_0$ yields:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq (1 - \eta\underline{\alpha})^k\mathbb{E}\|x^0 - x^*\|^2 + \frac{1}{\eta}(\underline{\alpha}\Delta_0 + \Delta_2), \quad \forall k < k_0 \tag{45}$$

For all $k \geq k_0$, (42) becomes:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq (1 - \frac{\eta\overline{\alpha}}{k+1})\mathbb{E}\|x^k - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{(k+1)^2} + \frac{\overline{\alpha}\Delta_2}{k+1}, \quad \forall k \geq k_0. \tag{46}$$

Note that $1 - \frac{\eta\overline{\alpha}}{k+1} \in (0, \frac{3}{4})$ when $k \geq k_0$ because $\frac{\eta\overline{\alpha}}{k+1} \leq \frac{\eta\overline{\alpha}}{k_0+1} \leq \eta\underline{\alpha} < \frac{1}{4}$. We use induction to prove that:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \frac{\Delta_1}{k+1} + \overline{\alpha}\Delta_2, \quad \forall k \geq k_0 \tag{47}$$

where

$$\Delta_1 = \max\left\{\frac{\overline{\alpha}^2\Delta_0}{\eta\overline{\alpha} - 1}, (k_0+1)\mathbb{E}\|x^{k_0} - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{k_0+1}\right\}.$$

When $k = k_0$, (47) holds because by (46) it follows:

$$\begin{aligned}
\mathbb{E}\|x^{k_0+1} - x^*\|^2 &\leq (1 - \frac{\eta\overline{\alpha}}{k_0+1})\mathbb{E}\|x^{k_0} - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{(k_0+1)^2} + \frac{\overline{\alpha}\Delta_2}{k_0+1} \\
&\leq \mathbb{E}\|x^{k_0} - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{(k_0+1)^2} + \overline{\alpha}\Delta_2 \\
&\leq \frac{\Delta_1}{k_0+1} + \overline{\alpha}\Delta_2.
\end{aligned} \tag{48}$$

Then, we assume that (47) holds for a certain $k \geq k_0$ and establish an upper bound for $\mathbb{E}\|x^{k+2} - x^*\|^2$ as:

$$\begin{aligned}
\mathbb{E}\|x^{k+2} - x^*\|^2 \leq& (1 - \frac{\eta\overline{\alpha}}{k+2})\mathbb{E}\|x^{k+1} - x^*\|^2 + \frac{\overline{\alpha}^2\Delta_0}{(k+2)^2} + \frac{\overline{\alpha}\Delta_2}{k+2} \\
\leq& (1 - \frac{\eta\overline{\alpha}}{k+2})\frac{\Delta_1}{k+1} + \frac{\overline{\alpha}^2\Delta_0}{(k+2)^2} + \overline{\alpha}\Delta_2 + (1 - \eta\overline{\alpha})\frac{\Delta_2}{k+2} \\
\overset{(e)}{\leq}& (1 - \frac{\eta\overline{\alpha}}{k+2})\frac{\Delta_1}{k+1} + \frac{\overline{\alpha}^2\Delta_0}{(k+2)^2} + \overline{\alpha}\Delta_2 \\
\overset{(f)}{\leq}& (1 - \frac{\eta\overline{\alpha}}{k+2})\frac{\Delta_1}{k+1} + \frac{(\eta\overline{\alpha} - 1)\Delta_1}{(k+2)^2} + \overline{\alpha}\Delta_2 \\
\leq& (1 - \frac{\eta\overline{\alpha}}{k+2})\frac{\Delta_1}{k+1} + \frac{(\eta\overline{\alpha} - 1)\Delta_1}{(k+1)(k+2)} + \overline{\alpha}\Delta_2 \\
\leq& \frac{1}{k+2}\Delta_1 + \overline{\alpha}\Delta_2
\end{aligned} \tag{49}$$

where (e) uses the fact that $\overline{\alpha} > \frac{1}{\eta}$, and (f) follows from $\Delta_1 \geq \frac{\overline{\alpha}^2\Delta_0}{\eta\overline{\alpha} - 1}$. This completes the induction as well as the proof.

Convergence of RSA with $\mathcal{O}(1/\sqrt{k})$ Step Size

Define the objective function in (9) as:

$$h_p(x) := \sum_{i\in\mathcal{R}} \Big(\mathbb{E}[F(x_i, \xi_i)] + \lambda||x_i - x_0||_p\Big) + f_0(x_0). \tag{50}$$

We have the following theorem for RSA with $\mathcal{O}(1/\sqrt{k})$ step size.

Theorem 4. *Suppose that Assumptions 1, 2 and 3 hold. Set the step size of ℓ_p-norm RSA ($p \geq 1$) as $\alpha^{k+1} = \min\{\underline{\alpha}, \frac{\overline{\alpha}}{\sqrt{k+1}}\}$, where $\underline{\alpha} = \min\{\min_{i\in\mathcal{R}} \frac{\mu_i}{4L_i^2}, \frac{\mu_0-\epsilon}{4L_0^2}\}$, $\overline{\alpha} > 0$, and ϵ is any constant within $(0, \mu_0)$, then we have:*

$$\mathbb{E}\Big[h_p(\bar{x}^k) - h_p(x^*)\Big] \leq \frac{\mathbb{E}\|x^0 - x^*\|^2 + \Delta_0 \sum_{\tau=0}^{k}(\alpha^{\tau+1})^2}{2\sum_{\tau=0}^{k}\alpha^{\tau+1}} + \frac{\Delta_2}{2} = \mathcal{O}\Big(\frac{\log k}{\sqrt{k}}\Big) + \frac{\Delta_2}{2} \tag{51}$$

where $\bar{x}^k$ is the running average solution

$$\bar{x}^k = \frac{\sum_{\tau=0}^{k}\alpha^{\tau+1}x^\tau}{\sum_{\tau=0}^{k}\alpha^{\tau+1}},$$

while Δ_0 and $\Delta_2 = \mathcal{O}(\lambda^2 q^2)$ are constants defined as

$$\Delta_0 = 16\lambda^2 rd + 16\lambda^2 r^2 d + 2\lambda^2 q^2 d + 2\sum_{i\in\mathcal{R}}\delta_i^2 \quad \textit{and} \quad \Delta_2 = \frac{\lambda^2 q^2 d}{\epsilon}.$$

Proof. For those equalities and inequalities that also appear in the proof of Theorem 2, we shall directly cite them. The proof contains the following steps.

Step 1. From the RSA update (10) at every regular worker i, corresponding to (28), we have:

$$\begin{aligned}
\mathbb{E}\|x_i^{k+1} - x_i^*\|^2 =& \mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
&- 2\alpha^{k+1}\mathbb{E}\langle\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big], x_i^k - x_i^*\rangle - 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle \\
\overset{(a)}{\leq}& \mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
&- 2\alpha^{k+1}\mathbb{E}(\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \mathbb{E}\big[F(x_i^*, \xi_i^k)\big] + \frac{\mu_i}{2}\|x_i^k - x_i^*\|^2) \\
&- 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle
\end{aligned} \tag{52}$$

where (a) is due to the strong convexity of $\mathbb{E}\big[F(x_i, \xi_i^k)\big]$.

For the second term at the right-hand side of (52), we have:

$$\begin{aligned}
&\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 \\
=&\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial\|x_i^k - x_0^k\|_p + \nabla F(x_i^k, \xi_i^k) - \nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big]\|^2 \\
\leq&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 + 2\|\nabla F(x_i^k, \xi_i^k) - \nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big]\|^2 \\
\overset{(b)}{\leq}&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2 + 2\delta_i^2
\end{aligned} \tag{53}$$

where (b) is due to the bounded variance given by Assumption 3. Plugging the optimality condition $\mathbb{E}\big[\nabla F(x_i^*, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^* - x_0^*\|_p = 0$ into $\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] + \lambda\partial_{x_i}\|x_i^k - x_0^k\|_p\|^2$ in the right-hand side of (53) yields:

$$\begin{aligned}
&\mathbb{E}\|\nabla F(x_i^k, \xi_i^k) + \lambda\partial\|x_i^k - x_0^k\|_p\|^2 \\
\leq&2\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \mathbb{E}\big[F(x_i^*, \xi_i^k)\big] + \lambda\partial\|x_i^k - x_0^k\|_p - \lambda\partial\|x_i^* - x_0^*\|_p\|^2 + 2\delta_i^2 \\
\leq&4\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 + 4\lambda^2\mathbb{E}\|\partial\|x_i^k - x_0^k\|_p - \partial\|x_i^* - x_0^*\|_p\|^2 + 2\delta_i^2 \\
\overset{(c)}{\leq}&4\mathbb{E}\|\nabla\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \nabla\mathbb{E}\big[F(x_i^*, \xi_i^k)\big]\|^2 + 16\lambda^2 d + 2\delta_i^2 \\
\overset{(d)}{\leq}&4L_i^2\mathbb{E}\|x_i^k - x_i^*\|^2 + 16\lambda^2 d + 2\delta_i^2
\end{aligned} \tag{54}$$

where (c) holds true because the b-norm of $\partial_{x_i}\|x_i^k - x_0^k\|_p$ (or $\partial_{x_i}\|x_i^* - x_0^*\|_p$) is no larger than 1 according to Proposition 1, and thus the absolute value of every element of the d-dimensional vector $\partial_{x_i}\|x_i^k - x_0^k\|_p$ (or $\partial_{x_i}\|x_i^* - x_0^*\|_p$) is no larger than 1, and (d) is due to the Lipschitz continuous gradients of $F(x_i, \xi_i^k)$. Substituting (54) into (52), we have:

$$\begin{aligned}
\mathbb{E}\|x_i^{k+1} - x_i^*\|^2 \leq&\mathbb{E}\|x_i^k - x_i^*\|^2 - 2\alpha^{k+1}\mathbb{E}(\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \mathbb{E}\big[F(x_i^*, \xi_i^k)\big]) \\
&- \alpha^{k+1}(\mu_i - 4\alpha^{k+1}L_i^2)\mathbb{E}\|x_i^k - x_i^*\|^2 + (\alpha^{k+1})^2(16\lambda^2 d + 2\delta_i^2) \\
&- 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle \\
\leq&\mathbb{E}\|x_i^k - x_i^*\|^2 - 2\alpha^{k+1}(\mathbb{E}\big[F(x_i^k, \xi_i^k)\big] - \mathbb{E}\big[F(x_i^*, \xi_i^k)\big]) \\
&- 2\alpha^{k+1}\mathbb{E}\langle\lambda\partial_{x_i}\|x_i^k - x_0^k\|_p, x_i^k - x_i^*\rangle + (\alpha^{k+1})^2(16\lambda^2 d + 2\delta_i^2)
\end{aligned} \tag{55}$$

We drop the term of $\mathbb{E}\|x_i^k - x_i^*\|^2$ because $\mu_i - 4\alpha^{k+1}L_i^2 \geq 0$ according to the step size rule.

Step 2. From the RSA update (11) at the master, we have:

$$\begin{aligned}
&\mathbb{E}\|x_0^{k+1} - x_0^*\|^2 \\
=&\mathbb{E}\left\|x_0^k - x_0^* - \alpha^{k+1}\left(\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right)\right\|^2 \\
=&\mathbb{E}\|x_0^k - x_0^*\|^2 + (\alpha^{k+1})^2\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\|^2 \\
&- 2\alpha^{k+1}\mathbb{E}\left\langle\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p, x_0^k - x_0^*\right\rangle - 2\alpha^{k+1}\mathbb{E}\langle\lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p, x_0^k - x_0^*\rangle
\end{aligned} \tag{56}$$

which is the same as (33).

For the second term at the right-hand side of (56), corresponding (57), we have:

$$\begin{aligned}
&\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\|^2 \\
\leq&2\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\|^2 + 2\lambda^2\mathbb{E}\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\|^2
\end{aligned} \tag{57}$$

For $\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\|^2$, we insert the optimality condition of (9) with respect to x_0, namely, $\nabla f_0(x_0^*) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_i^* - x_0^*\|_p = 0$ to obtain:

$$\begin{aligned}
&\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p\|^2 \\
\leq&\mathbb{E}\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p - \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_i^* - x_0^*\|_p\|^2 \\
\leq&2\mathbb{E}\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*)\|^2 + 2\lambda^2\mathbb{E}\|\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p - \sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^* - x_i^*\|_p\|^2 \\
\leq&2\mathbb{E}\|\nabla f_0(x_0^k) - \nabla f_0(x_0^*)\|^2 + 8\lambda^2r^2d \\
\overset{(e)}{\leq}&2L_0^2\mathbb{E}\|x_0^k - x_0^*\|^2 + 8\lambda^2r^2d
\end{aligned} \tag{58}$$

where (e) is due to the Lipschitz continuous gradients of $f_0(x_0)$. Substituting $\mathbb{E}\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\|^2 \leq q^2d$ in (35) and (58) into (57) yields:

$$\mathbb{E}\|\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p + \lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\|^2 \leq 4L_0^2\mathbb{E}\|x_0^k - x_0^*\|^2 + 16\lambda^2r^2d + 2\lambda^2q^2d \tag{59}$$

For the third term at the right-hand side of (56), since $f_0(x_0)$ is strongly convex with constant μ_0, we have:

$$\begin{aligned}
&\mathbb{E}\langle\nabla f_0(x_0^k) + \lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p, x_0^k - x_0^*\rangle \\
\geq&\mathbb{E}(f_0(x_0^k) - f_0(x_0^*) + \frac{\mu_0}{2}\|x_0^k - x_0^*\|^2) + \mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k - x_i^k\|_p, x_0^k - x_0^*\rangle
\end{aligned} \tag{60}$$

For the last term at the right-hand side of (56), it holds for any $\epsilon > 0$ that:

$$\begin{aligned}
2\mathbb{E}\left\langle\lambda\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p, x_0^k - x_0^*\right\rangle \leq&\epsilon\mathbb{E}\|x_0^k - x_0^*\|^2 + \frac{\lambda^2}{\epsilon}\mathbb{E}\left\|\sum_{j\in\mathcal{B}}\partial_{x_0}\|x_0^k - z_j^k\|_p\right\|^2 \\
\leq&\epsilon\mathbb{E}\|x_0^k - x_0^*\|^2 + \frac{\lambda^2q^2d}{\epsilon}.
\end{aligned} \tag{61}$$

Substituting (59), (60) and (61) into (56), we have:

$$\begin{aligned}\mathbb{E}\|x_0^{k+1}-x_0^*\|^2 \leq & \mathbb{E}\|x_0^k-x_0^*\|^2+(\alpha^{k+1})^2(16\lambda^2r^2d+2\lambda^2q^2d)+\frac{\alpha^{k+1}\lambda^2q^2d}{\epsilon}\\ & -2\alpha^{k+1}\mathbb{E}[f_0(x_0^k)-f_0(x_0^*)]-2\alpha^{k+1}\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p,x_0^k-x_0^*\rangle\\ & -\alpha^{k+1}(\mu_0-4L_0^2\alpha^{k+1}-\epsilon)\mathbb{E}\|x_0^k-x_0^*\|^2 \qquad (62)\\ \leq & \mathbb{E}\|x_0^k-x_0^*\|^2+(\alpha^{k+1})^2(16\lambda^2r^2d+2\lambda^2q^2d)+\frac{\alpha^{k+1}\lambda^2q^2d}{\epsilon}\\ & -2\alpha^{k+1}\mathbb{E}[f_0(x_0^k)-f_0(x_0^*)]-2\alpha^{k+1}\mathbb{E}\langle\lambda\sum_{i\in\mathcal{R}}\partial_{x_0}\|x_0^k-x_i^k\|_p,x_0^k-x_0^*\rangle\end{aligned}$$

where we drop the term of $\mathbb{E}\|x_0^k-x_0^*\|^2$ because $\mu_0-4L_0^2\alpha^{k+1}-\epsilon\geq 0$ according to the step size rule.
Step 3. Using the convexity of $\|x_i-x_0\|_p$, we have:

$$\sum_{i\in\mathcal{R}}\langle\partial_{x_i}\|x_0^k-x_i^k\|_p,x_i^k-x_i^*\rangle+\sum_{i\in\mathcal{R}}\langle\partial_{x_0}\|x_0^k-x_i^k\|_p,x_0^k-x_0^*\rangle\geq\sum_{i\in\mathcal{R}}\|x_i^k-x_0^k\|_p-\sum_{i\in\mathcal{R}}\|x_i^*-x_0^*\|_p. \qquad (63)$$

Summing up (55) for all $i\in\mathcal{R}$, as well as combining (62) and (63), we have:

$$2\alpha^{k+1}\mathbb{E}(h_p(x^k)-h_p(x^*))\leq\mathbb{E}\|x^k-x^*\|^2-\mathbb{E}\|x^{k+1}-x^*\|^2+(\alpha^{k+1})^2\Delta_0+\alpha^{k+1}\Delta_2. \qquad (64)$$

where Δ_0 and Δ_2 are constants defined in (43). Summing up (64) for all times k, we have:

$$2\sum_{\tau=0}^{k}\alpha^{\tau+1}E\left[\sum_{\tau=0}^{k}\frac{\alpha^{\tau+1}}{\sum_{\tau=0}^{k}\alpha^{\tau+1}}(h_p(x^\tau)-h_p(x^*))\right]\leq\mathbb{E}\|x^0-x^*\|^2+\Delta_0\sum_{\tau=0}^{k}(\alpha^{\tau+1})^2+\Delta_2\sum_{\tau=0}^{k}\alpha^{\tau+1} \qquad (65)$$

Since $h_p(x)$ is convex, we have:

$$\sum_{\tau=0}^{k}\frac{\alpha^{\tau+1}}{\sum_{\tau=0}^{k}\alpha^{\tau+1}}(h_p(x^\tau)-h_p(x^*))\geq h_p(\bar{x}^k)-h_p(x^*). \qquad (66)$$

Substituting (66) to (65) yields:

$$\mathbb{E}\left[h_p(\bar{x}^k)-h_p(x^*)\right]\leq\frac{\mathbb{E}\|x^0-x^*\|^2+\Delta_0\sum_{\tau=0}^{k}(\alpha^{\tau+1})^2}{2\sum_{\tau=0}^{k}\alpha^{\tau+1}}+\frac{\Delta_2}{2}. \qquad (67)$$

Proof of Theorem 3

Proof. When $\lambda\geq\lambda_0$, combining Theorem 1 and Theorem 2 directly yields (14). When $0<\lambda<\lambda_0$, we have:

$$\mathbb{E}[\|x^k-[\tilde{x}^*]\|^2]\leq 2\mathbb{E}[\|x^k-x^*\|^2]+2\mathbb{E}[\|x^*-[\tilde{x}^*]\|^2]$$

where the inequality follows from $(a+b)^2\leq 2a^2+2b^2$. By Theorem 2 and $\mathbb{E}[\|x^*-[\tilde{x}^*]\|^2]\leq\Delta_3$, (15) holds true.